Biostat 200C Homework 1

Due Apr 16 @ 11:59PM

Zian ZHUANG

Q1. Binomial Distribution

Let Y_i be the number of successes in n_i trials with

$$Y_i \sim Bin(n_i, \pi_i),$$

where the probabilities π_i have a Beta distribution

$$\pi_i \sim Beta(\alpha, \beta).$$

The probability density function for the Beta distribution is $f(x; \alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha, \beta)$ for $x \in [0, 1], \alpha > 0, \beta > 0$, and the beta function $B(\alpha, \beta)$ defining the normalizing constant required to ensure that $\int_0^1 f(x; \alpha, \beta) = 1$. Let $\theta = \alpha/(\alpha + \beta)$, show that

a. $E(\pi_i) = \theta$

$$\begin{aligned}
E(\pi_i) &= \int \pi_i * f(\pi_i) d\pi_i \\
&= \int \pi_i * \pi_i^{\alpha-1}(1-\pi_i)^{\beta-1}/B(\alpha, \beta) d\pi_i \\
&= B(\alpha, \beta)^{-1} \int \pi_i^{(\alpha+1)-1}(1-\pi_i)^{\beta-1} d\pi_i \\
&= B(\alpha+1, \beta) * B(\alpha, \beta)^{-1} \int B(\alpha+1, \beta)^{-1} * \pi_i^{(\alpha+1)-1}(1-\pi_i)^{\beta-1} d\pi_i \\
&= B(\alpha+1, \beta) * B(\alpha, \beta)^{-1} * 1 \\
&= \Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+1+\beta)/(\Gamma(\alpha)\Gamma(\beta)) * \Gamma(\alpha+\beta) \\
&= \alpha/(\alpha+\beta) \\
&= \theta
\end{aligned}$$

b. $Var(\pi_i) = \theta(1-\theta)/(\alpha+\beta+1) = \phi\theta(1-\theta)$ Firstly we can calculated $E(\pi_i^2)$

$$\begin{aligned}
E(\pi_i^2) &= \int \pi_i^2 * f(\pi_i) d\pi_i \\
&= \int \pi_i^2 * \pi_i^{\alpha-1}(1-\pi_i)^{\beta-1}/B(\alpha,\beta) d\pi_i \\
&= B(\alpha,\beta)^{-1} \int \pi_i^{(\alpha+2)-1}(1-\pi_i)^{\beta-1} d\pi_i \\
&= B(\alpha+2,\beta) * B(\alpha,\beta)^{-1} \int B(\alpha+1,\beta)^{-1} * \pi_i^{(\alpha+2)-1}(1-\pi_i)^{\beta-1} d\pi_i \\
&= B(\alpha+2,\beta) * B(\alpha,\beta)^{-1} * 1 \\
&= \Gamma(\alpha+2)\Gamma(\beta)/\Gamma(\alpha+2+\beta)/(\Gamma(\alpha)\Gamma(\beta)) * \Gamma(\alpha+\beta) \\
&= \alpha * (\alpha+1)/(\alpha+1+\beta) * (\alpha+\beta) \\
&= \theta(\alpha+1)/(\alpha+1+\beta)
\end{aligned}$$

Then we can obtain $Var(\pi_i)$

$$\begin{aligned}
Var(\pi_i) &= E(\pi_i^2) - E(\pi_i)^2 \\
&= ((\alpha+1)\alpha(\alpha+\beta) - \alpha^2(\alpha+\beta+1))/(\alpha+\beta+1)(\alpha+\beta)^2 \\
&= (\alpha\beta)/(\alpha+\beta)^2/(\alpha+1+\beta) \\
&= \theta(1-\theta)/(\alpha+\beta+1) = \phi\theta(1-\theta)
\end{aligned}$$

c. $E(Y_i) = n_i\theta$

$$\begin{aligned}
E(Y_i) &= E_{\pi_i}(E_{Y_i}(Y_i|\pi_i)) \\
&= E_{\pi_i}(n_i * \pi_i) \\
&= n_i * E(\pi_i) \\
&= n_i * \theta
\end{aligned}$$

d. $Var(Y_i) = n_i\theta(1-\theta)[1+(n_i-1)\phi]$ so that $Var(Y_i)$ is larger than the Binomial variance (unless $n_i = 1$ or $\phi = 0$).

$$\begin{aligned}
Var(Y_i) &= E_{\pi_i}(Var(Y_i|\pi_i)) + Var_{\pi_i}(E(Y_i|\pi_i)) \\
&= E_{\pi_i}(n_i * \pi_i * (1-\pi_i)) + Var_{\pi_i}(\pi_i * n_i) \\
&= n_i * (E(\pi_i) - E(\pi_i^2)) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i * (\theta - \theta(\alpha+1)/(\alpha+1+\beta)) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i * (\theta(1 - (\alpha+1)/(\alpha+1+\beta))) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i * (\theta * \beta/(\alpha+1+\beta)) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i * (\theta * (1-\theta)(1-\phi)) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i\theta(1-\theta)[1+(n_i-1)\phi]
\end{aligned}$$

Q2. (ELMR Chapter 3 Exercise 1)

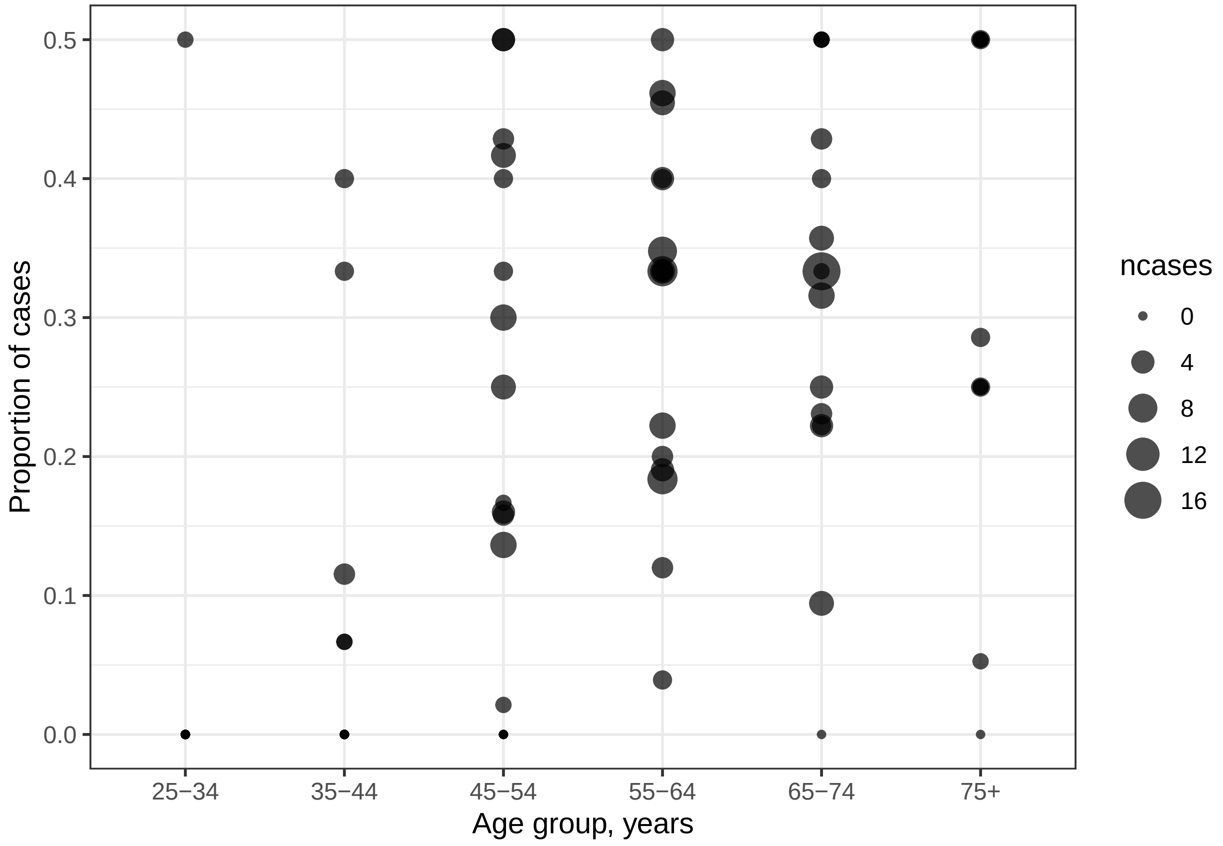
A case-control study of esophageal cancer in Ileet-Vilaine, France.

```
data(esoph)
#help(esoph)
```

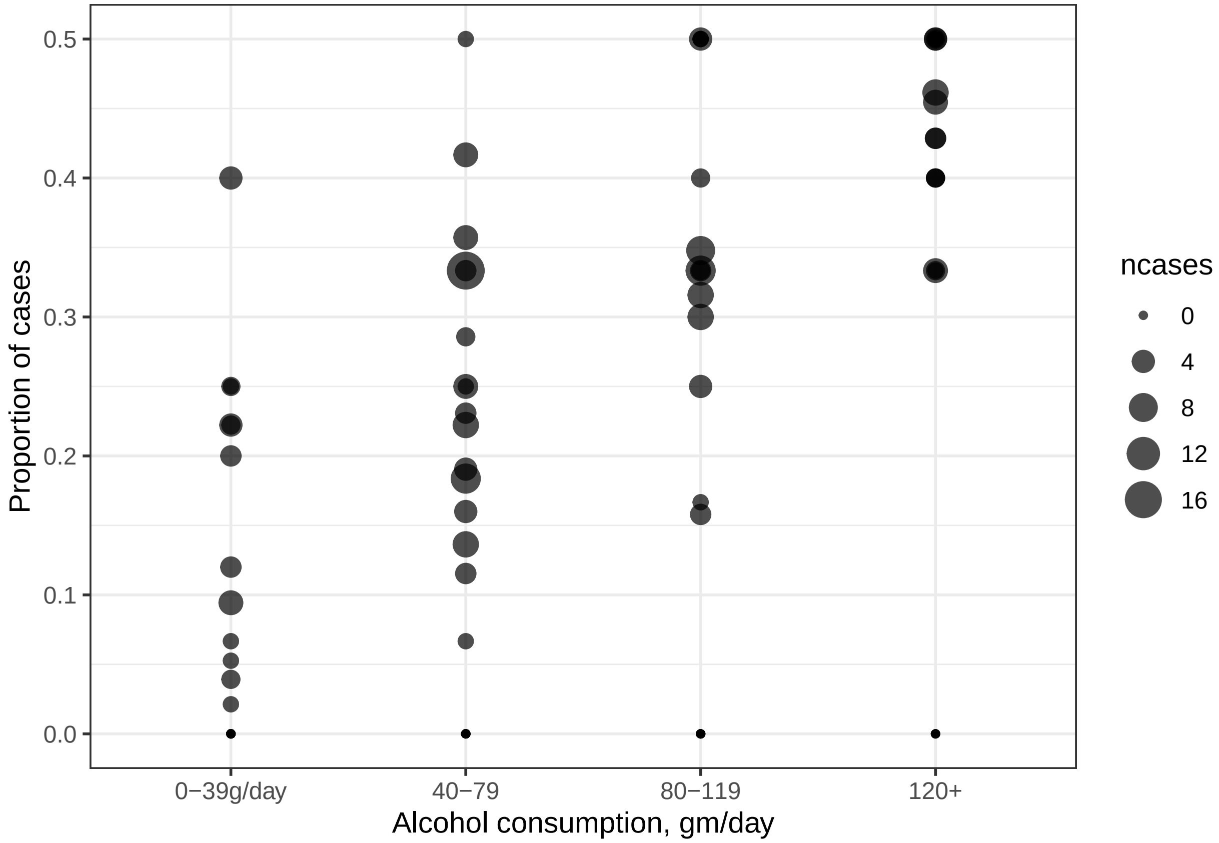
a. Plot the proportion of cases against each predictor using the size of the point to indicate the number of subject as seen in Figure 2.7. Comment on the realtionships seen in the plots.

Solution:

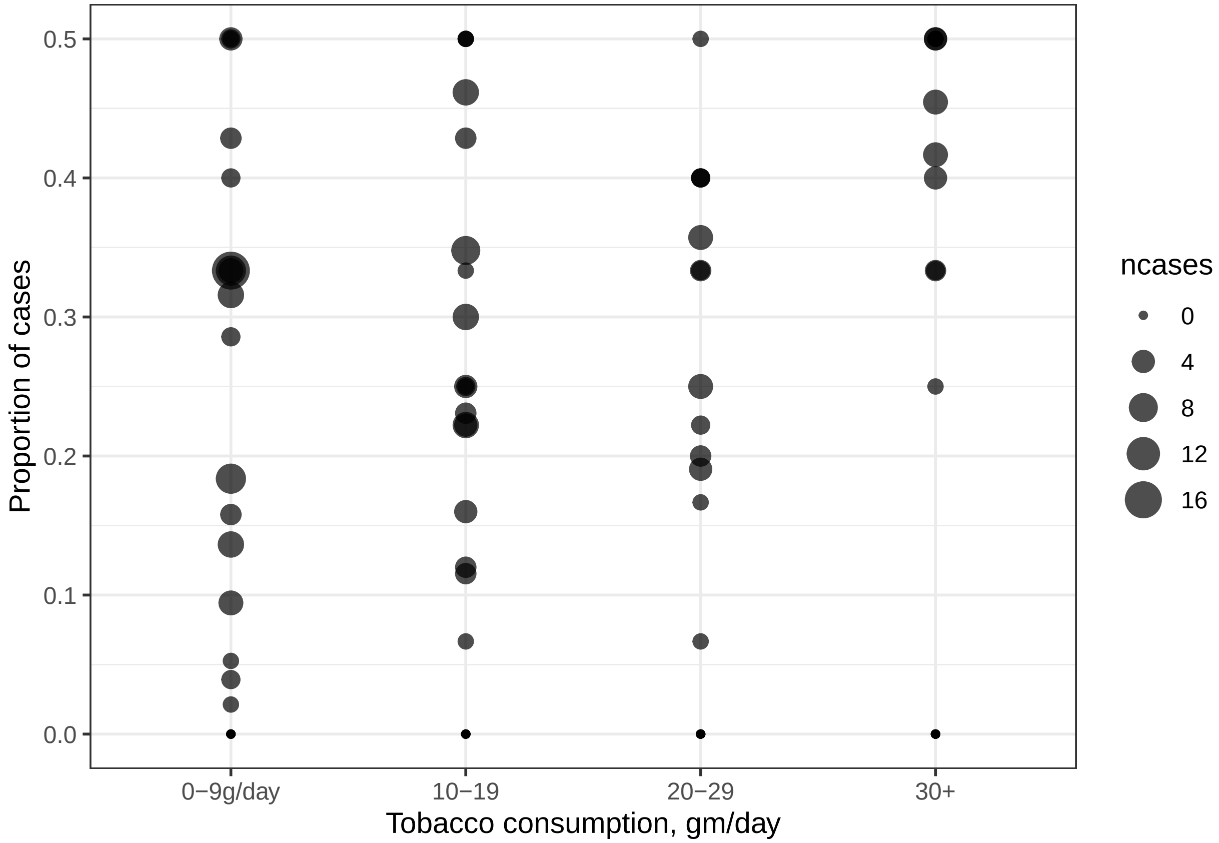
```
plot_data <- esoph %>%
  mutate(proportion=ncases/(ncontrols+ncases))

ggplot(plot_data, aes(agegp, proportion))+
  geom_point(aes(size = ncases),alpha = 7/10)+
  ylab("Proportion of cases")+
  xlab("Age group, years")+
  theme_bw()
```



```
ggplot(plot_data, aes(alcgp, proportion))+
  geom_point(aes(size = ncases),alpha = 7/10)+
  ylab("Proportion of cases")+
  xlab("Alcohol consumption, gm/day")+
  theme_bw()
```



```
ggplot(plot_data, aes(tobgp, proportion))+
  geom_point(aes(size = ncases),alpha = 7/10)+
  ylab("Proportion of cases")+
  xlab("Tobacco consumption, gm/day")+
  theme_bw()
```



b. Fit a binomial GLM with interactions between all three predictors. Use AIC as a criterion to select a model using the `step` function. Which model is selected?

Solution:

```
lmod = glm(cbind(ncases, ncontrols)~agegp*alcgp*tobgp,
           family = binomial, data=esoph)
summary(lmod)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp * alcgp * tobgp,
##     family = binomial, data = esoph)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [26]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [51]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [76]  0  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients: (8 not defined because of singularities)
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.647e+01  4.487e+05       0        1
## agegp.L                 -1.557e+01  1.619e+06       0        1
## agegp.Q                 -4.025e+01  1.484e+06       0        1
## agegp.C                 -1.468e+01  1.009e+06       0        1
## agegp^4                 -1.230e+01  5.474e+05       0        1
## agegp^5                  1.268e-01  1.956e+05       0        1
## alcgp.L                 -2.383e+01  9.105e+05       0        1
## alcgp.Q                 -1.367e+01  5.379e+05       0        1
## alcgp.C                 -7.679e+00  4.468e+05       0        1
## tobgp.L                 -1.955e+01  9.947e+05       0        1
## tobgp.Q                 -1.309e+01  5.084e+05       0        1
## tobgp.C                 -3.142e+00  2.729e+05       0        1
## agegp.L:alcgp.L         -1.394e+02  3.511e+06       0        1
## agegp.Q:alcgp.L         -1.211e+02  3.133e+06       0        1
## agegp.C:alcgp.L         -5.901e+01  2.002e+06       0        1
## agegp^4:alcgp.L         -5.601e+01  1.313e+06       0        1
## agegp^5:alcgp.L         -1.224e+01  3.374e+05       0        1
## agegp.L:alcgp.Q         -7.314e+01  2.344e+06       0        1
## agegp.Q:alcgp.Q         -5.221e+01  2.046e+06       0        1
## agegp.C:alcgp.Q         -2.645e+01  1.141e+06       0        1
## agegp^4:alcgp.Q         -3.742e+01  9.698e+05       0        1
## agegp^5:alcgp.Q          9.039e+00  1.491e+05       0        1
## agegp.L:alcgp.C         -5.293e+01  1.799e+06       0        1
## agegp.Q:alcgp.C         -4.331e+01  1.600e+06       0        1
## agegp.C:alcgp.C         -1.528e+01  9.243e+05       0        1
## agegp^4:alcgp.C         -2.349e+01  6.696e+05       0        1
## agegp^5:alcgp.C         -1.180e+00  1.099e+05       0        1
## agegp.L:tobgp.L         -9.389e+01  3.715e+06       0        1
## agegp.Q:tobgp.L         -7.727e+01  3.347e+06       0        1
## agegp.C:tobgp.L         -3.353e+01  2.229e+06       0        1
## agegp^4:tobgp.L         -4.109e+01  1.341e+06       0        1
## agegp^5:tobgp.L          5.377e+00  4.011e+05       0        1
## agegp.L:tobgp.Q         -5.881e+01  2.177e+06       0        1
## agegp.Q:tobgp.Q         -5.316e+01  1.890e+06       0        1
## agegp.C:tobgp.Q         -2.083e+01  1.094e+06       0        1
## agegp^4:tobgp.Q         -2.220e+01  9.134e+05       0        1
## agegp^5:tobgp.Q         -1.313e+00  1.626e+05       0        1
## agegp.L:tobgp.C         -2.543e+01  1.210e+06       0        1
## agegp.Q:tobgp.C         -1.423e+01  1.051e+06       0        1
## agegp.C:tobgp.C         -7.456e+00  5.331e+05       0        1
## agegp^4:tobgp.C         -9.394e+00  5.386e+05       0        1
## agegp^5:tobgp.C          2.459e+00  4.054e+04       0        1
## alcgp.L:tobgp.L         -5.552e+01  2.111e+06       0        1
## alcgp.Q:tobgp.L         -4.120e+01  1.271e+06       0        1
## alcgp.C:tobgp.L         -2.234e+01  9.951e+05       0        1
## alcgp.L:tobgp.Q         -1.053e+01  1.190e+06       0        1
## alcgp.Q:tobgp.Q         -3.193e+01  7.778e+05       0        1
## alcgp.C:tobgp.Q         -1.231e+01  4.155e+05       0        1
## alcgp.L:tobgp.C          1.387e+01  4.514e+05       0        1
## alcgp.Q:tobgp.C         -1.718e+01  3.995e+05       0        1
## alcgp.C:tobgp.C          4.239e-01  2.647e+04       0        1
## agegp.L:alcgp.L:tobgp.L -2.840e+02  8.518e+06       0        1
## agegp.Q:alcgp.L:tobgp.L -2.551e+02  7.548e+06       0        1
## agegp.C:alcgp.L:tobgp.L -1.191e+02  4.618e+06       0        1
## agegp^4:alcgp.L:tobgp.L -1.355e+02  3.302e+06       0        1
## agegp^5:alcgp.L:tobgp.L -1.160e+01  8.210e+05       0        1
## agegp.L:alcgp.Q:tobgp.L -1.933e+02  5.800e+06       0        1
## agegp.Q:alcgp.Q:tobgp.L -1.619e+02  5.002e+06       0        1
## agegp.C:alcgp.Q:tobgp.L -7.717e+01  2.661e+06       0        1
## agegp^4:alcgp.Q:tobgp.L -8.845e+01  2.509e+06       0        1
## agegp^5:alcgp.Q:tobgp.L  7.564e+00  2.634e+05       0        1
## agegp.L:alcgp.C:tobgp.L -1.086e+02  4.072e+06       0        1
## agegp.Q:alcgp.C:tobgp.L -9.399e+01  3.573e+06       0        1
## agegp.C:alcgp.C:tobgp.L -4.413e+01  2.055e+06       0        1
## agegp^4:alcgp.C:tobgp.L -4.470e+01  1.524e+06       0        1
## agegp^5:alcgp.C:tobgp.L -3.678e+00  2.280e+05       0        1
## agegp.L:alcgp.L:tobgp.Q -1.033e+02  5.251e+06       0        1
## agegp.Q:alcgp.L:tobgp.Q -8.689e+01  4.573e+06       0        1
## agegp.C:alcgp.L:tobgp.Q -1.386e+01  2.546e+06       0        1
## agegp^4:alcgp.L:tobgp.Q -8.360e+01  2.212e+06       0        1
## agegp^5:alcgp.L:tobgp.Q  1.969e+01  4.519e+05       0        1
## agegp.L:alcgp.Q:tobgp.Q -1.631e+02  3.877e+06       0        1
## agegp.Q:alcgp.Q:tobgp.Q -1.447e+02  3.275e+06       0        1
## agegp.C:alcgp.Q:tobgp.Q -5.394e+01  1.606e+06       0        1
## agegp^4:alcgp.Q:tobgp.Q -7.473e+01  1.773e+06       0        1
## agegp^5:alcgp.Q:tobgp.Q         NA         NA      NA       NA
## agegp.L:alcgp.C:tobgp.Q -4.418e+01  1.984e+06       0        1
## agegp.Q:alcgp.C:tobgp.Q -4.237e+01  1.646e+06       0        1
## agegp.C:alcgp.C:tobgp.Q -2.202e+01  8.027e+05       0        1
## agegp^4:alcgp.C:tobgp.Q -1.989e+01  8.172e+05       0        1
## agegp^5:alcgp.C:tobgp.Q         NA         NA      NA       NA
## agegp.L:alcgp.L:tobgp.C  2.881e+01  2.178e+06       0        1
## agegp.Q:alcgp.L:tobgp.C  3.731e+01  1.871e+06       0        1
## agegp.C:alcgp.L:tobgp.C  1.822e+01  9.043e+05       0        1
## agegp^4:alcgp.L:tobgp.C  1.169e+01  9.950e+05       0        1
## agegp^5:alcgp.L:tobgp.C         NA         NA      NA       NA
## agegp.L:alcgp.Q:tobgp.C -8.620e+01  1.832e+06       0        1
## agegp.Q:alcgp.Q:tobgp.C -6.466e+01  1.511e+06       0        1
## agegp.C:alcgp.Q:tobgp.C -3.978e+01  8.913e+05       0        1
## agegp^4:alcgp.Q:tobgp.C -3.334e+01  8.036e+05       0        1
## agegp^5:alcgp.Q:tobgp.C         NA         NA      NA       NA
## agegp.L:alcgp.C:tobgp.C -2.805e+00  2.214e+05       0        1
## agegp.Q:alcgp.C:tobgp.C         NA         NA      NA       NA
## agegp.C:alcgp.C:tobgp.C         NA         NA      NA       NA
## agegp^4:alcgp.C:tobgp.C         NA         NA      NA       NA
## agegp^5:alcgp.C:tobgp.C         NA         NA      NA       NA
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.2724e+02  on 87  degrees of freedom
## Residual deviance: 3.0119e-10  on  0  degrees of freedom
## AIC: 323.48
##
## Number of Fisher Scoring iterations: 25
```

```
lmods = step(lmod, direction = "both")
```

```
## Start:  AIC=323.48
## cbind(ncases, ncontrols) ~ agegp * alcgp * tobgp
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##                      Df Deviance    AIC
## - agegp:alcgp:tobgp 37   16.109 265.59
## <none>                     0.000 323.48
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Step:  AIC=265.59
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     agegp:tobgp + alcgp:tobgp
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                      Df Deviance    AIC
## - agegp:tobgp       15   27.146 246.63
## - agegp:alcgp       15   34.364 253.84
## - alcgp:tobgp        9   23.776 255.26
## <none>                   16.109 265.59
## + agegp:alcgp:tobgp 37    0.000 323.48
##
## Step:  AIC=246.63
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     alcgp:tobgp
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                Df Deviance    AIC
## - alcgp:tobgp  9   33.796 235.28
## - agegp:alcgp 15   47.484 236.96
## <none>             27.146 246.63
## + agegp:tobgp 15   16.109 265.59
##
## Step:  AIC=235.28
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                Df Deviance    AIC
## - agegp:alcgp 15   53.973 225.45
## <none>             33.796 235.28
## - tobgp        3   44.151 239.63
## + alcgp:tobgp  9   27.146 246.63
## + agegp:tobgp 15   23.776 255.26
##
## Step:  AIC=225.45
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
##
##                Df Deviance    AIC
## <none>             53.973 225.45
## - tobgp        3   64.572 230.05
## + agegp:alcgp 15   33.796 235.28
## + alcgp:tobgp  9   47.484 236.96
## + agegp:tobgp 15   41.455 242.94
## - alcgp        3  120.028 285.51
## - agegp        5  131.484 292.96
```

Finally we selected `cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp` as the best mopdel according to the AIC criteria.

c. All three factors are ordered and so special contrasts have been used appropriate for ordered factors involving linear, quadratic and cubic terms. Further simplification of the model may be possible by eliminating some of these terms. Use the `unclass` function to convert the factors to a numerical representation and check whether the model may be simplified.

Solution:

```
lmod = glm(cbind(ncases, ncontrols)~unclass(agegp)*unclass(alcgp)*unclass(tobgp),
           family = binomial, data=esoph)
summary(lmod)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ unclass(agegp) * unclass(alcgp) *
##     unclass(tobgp), family = binomial, data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9913  -0.7770  -0.3314   0.2674   2.0156
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                  -8.06144    1.60279  -5.030
## unclass(agegp)                                0.95321    0.36256   2.629
## unclass(alcgp)                                1.70106    0.62226   2.734
## unclass(tobgp)                                0.94378    0.65165   1.448
## unclass(agegp):unclass(alcgp)                -0.17364    0.14683  -1.183
## unclass(agegp):unclass(tobgp)                -0.07549    0.15496  -0.487
## unclass(alcgp):unclass(tobgp)                -0.25483    0.25722  -0.991
## unclass(agegp):unclass(alcgp):unclass(tobgp)  0.02564    0.06384   0.402
##                                              Pr(>|z|)
## (Intercept)                                  4.91e-07 ***
## unclass(agegp)                                0.00856 **
## unclass(alcgp)                                0.00626 **
## unclass(tobgp)                                0.14753
## unclass(agegp):unclass(alcgp)                 0.23698
## unclass(agegp):unclass(tobgp)                 0.62616
## unclass(alcgp):unclass(tobgp)                 0.32184
## unclass(agegp):unclass(alcgp):unclass(tobgp)  0.68802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 227.24  on 87  degrees of freedom
## Residual deviance:  67.57  on 80  degrees of freedom
## AIC: 231.05
##
## Number of Fisher Scoring iterations: 4
```

```
lmods = step(lmod, direction = "both")
```

```
## Start:  AIC=231.05
## cbind(ncases, ncontrols) ~ unclass(agegp) * unclass(alcgp) *
##     unclass(tobgp)
##
##                                                Df Deviance    AIC
## - unclass(agegp):unclass(alcgp):unclass(tobgp)  1   67.732 229.21
## <none>                                              67.570 231.05
##
## Step:  AIC=229.21
## cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp) + unclass(agegp):unclass(alcgp) + unclass(agegp):unclass(tobgp) +
##     unclass(alcgp):unclass(tobgp)
##
##                                                Df Deviance    AIC
## - unclass(agegp):unclass(tobgp)                 1   67.813 227.29
## <none>                                              67.732 229.21
## - unclass(agegp):unclass(alcgp)                 1   70.772 230.25
## + unclass(agegp):unclass(alcgp):unclass(tobgp)  1   67.570 231.05
## - unclass(alcgp):unclass(tobgp)                 1   71.911 231.39
##
## Step:  AIC=227.29
## cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp) + unclass(agegp):unclass(alcgp) + unclass(alcgp):unclass(tobgp)
##
##                                 Df Deviance    AIC
## <none>                               67.813 227.29
## - unclass(agegp):unclass(alcgp)  1   70.852 228.33
## + unclass(agegp):unclass(tobgp)  1   67.732 229.21
## - unclass(alcgp):unclass(tobgp)  1   71.913 229.39
```

```
drop1(lmods, test = c("Chisq"))
```

```
## Single term deletions
##
## Model:
## cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) +
##     unclass(tobgp) + unclass(agegp):unclass(alcgp) + unclass(alcgp):unclass(tobgp)
##                               Df Deviance    AIC    LRT Pr(>Chi)
## <none>                             67.813 227.29
## unclass(agegp):unclass(alcgp)  1   70.852 228.33 3.0383  0.08132 .
## unclass(alcgp):unclass(tobgp)  1   71.913 229.39 4.0995  0.04290 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We selected `cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(alcgp) + unclass(tobgp) + unclass(agegp):unclass(alcgp) + unclass(alcgp):unclass(tobgp)` as the best model. After drop1 test (chisq), we found that interaction of agegp and alcgp is not significant within the 95% confidence interval.

d. Use the summary output of the factor model to suggest a model that is slightly more complex than the linear model proposed in the previous question.

Solution:

```
lmod = glm(cbind(ncases, ncontrols)~agegp*alcgp*tobgp,
           family = binomial, data=esoph)
lmods = step(lmod, direction = "both")
```

```
## Start:  AIC=323.48
## cbind(ncases, ncontrols) ~ agegp * alcgp * tobgp

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                      Df Deviance    AIC
## - agegp:alcgp:tobgp 37   16.109 265.59
## <none>                    0.000 323.48

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=265.59
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     agegp:tobgp + alcgp:tobgp

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                      Df Deviance    AIC
## - agegp:tobgp       15   27.146 246.63
## - agegp:alcgp       15   34.364 253.84
## - alcgp:tobgp        9   23.776 255.26
## <none>                   16.109 265.59
## + agegp:alcgp:tobgp 37    0.000 323.48
##
## Step:  AIC=246.63
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     alcgp:tobgp

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                Df Deviance    AIC
## - alcgp:tobgp  9   33.796 235.28
## - agegp:alcgp 15   47.484 236.96
## <none>             27.146 246.63
## + agegp:tobgp 15   16.109 265.59
##
## Step:  AIC=235.28
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                 Df Deviance    AIC
## - agegp:alcgp 15   53.973 225.45
## <none>             33.796 235.28
## - tobgp        3   44.151 239.63
## + alcgp:tobgp  9   27.146 246.63
## + agegp:tobgp 15   23.776 255.26
##
## Step:  AIC=225.45
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
##
##                 Df Deviance    AIC
## <none>             53.973 225.45
## - tobgp        3   64.572 230.05
## + agegp:alcgp 15   33.796 235.28
## + alcgp:tobgp  9   47.484 236.96
## + agegp:tobgp 15   41.455 242.94
## - alcgp        3  120.028 285.51
## - agegp        5  131.484 292.96
```

```
add1(lmods, lmod, test = c("F"))
```

```
## Warning in add1.glm(lmods, lmod, test = c("F")): F test assumes quasibinomial
## family
```

```
## Single term additions
##
## Model:
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
##             Df Deviance    AIC F value   Pr(>F)
## <none>           53.973 225.45
## agegp:alcgp 15   33.796 235.28  2.4279 0.007722 **
## agegp:tobgp 15   41.455 242.94  1.2280 0.276666
## alcgp:tobgp  9   47.484 236.96  1.0174 0.435386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the add1 test (F), we found that interaction of agegp and alcgp is significant within the 95% confidence interval and may be added in the model.

e. Does your final model fit the data? Is the test you make accurate for this data?

Solution:

```
lmod = glm(cbind(ncases, ncontrols)~agegp*alcgp*tobgp,
           family = binomial, data=esoph)
summary(lmod)
```

```
##
## Call:
```

```
## glm(formula = cbind(ncases, ncontrols) ~ agegp * alcgp * tobgp,
##     family = binomial, data = esoph)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [26]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [51]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [76]  0  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients: (8 not defined because of singularities)
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.647e+01  4.487e+05       0        1
## agegp.L                 -1.557e+01  1.619e+06       0        1
## agegp.Q                 -4.025e+01  1.484e+06       0        1
## agegp.C                 -1.468e+01  1.009e+06       0        1
## agegp^4                 -1.230e+01  5.474e+05       0        1
## agegp^5                  1.268e-01  1.956e+05       0        1
## alcgp.L                 -2.383e+01  9.105e+05       0        1
## alcgp.Q                 -1.367e+01  5.379e+05       0        1
## alcgp.C                 -7.679e+00  4.468e+05       0        1
## tobgp.L                 -1.955e+01  9.947e+05       0        1
## tobgp.Q                 -1.309e+01  5.084e+05       0        1
## tobgp.C                 -3.142e+00  2.729e+05       0        1
## agegp.L:alcgp.L         -1.394e+02  3.511e+06       0        1
## agegp.Q:alcgp.L         -1.211e+02  3.133e+06       0        1
## agegp.C:alcgp.L         -5.901e+01  2.002e+06       0        1
## agegp^4:alcgp.L         -5.601e+01  1.313e+06       0        1
## agegp^5:alcgp.L         -1.224e+01  3.374e+05       0        1
## agegp.L:alcgp.Q         -7.314e+01  2.344e+06       0        1
## agegp.Q:alcgp.Q         -5.221e+01  2.046e+06       0        1
## agegp.C:alcgp.Q         -2.645e+01  1.141e+06       0        1
## agegp^4:alcgp.Q         -3.742e+01  9.698e+05       0        1
## agegp^5:alcgp.Q          9.039e+00  1.491e+05       0        1
## agegp.L:alcgp.C         -5.293e+01  1.799e+06       0        1
## agegp.Q:alcgp.C         -4.331e+01  1.600e+06       0        1
## agegp.C:alcgp.C         -1.528e+01  9.243e+05       0        1
## agegp^4:alcgp.C         -2.349e+01  6.696e+05       0        1
## agegp^5:alcgp.C         -1.180e+00  1.099e+05       0        1
## agegp.L:tobgp.L         -9.389e+01  3.715e+06       0        1
## agegp.Q:tobgp.L         -7.727e+01  3.347e+06       0        1
## agegp.C:tobgp.L         -3.353e+01  2.229e+06       0        1
## agegp^4:tobgp.L         -4.109e+01  1.341e+06       0        1
## agegp^5:tobgp.L          5.377e+00  4.011e+05       0        1
## agegp.L:tobgp.Q         -5.881e+01  2.177e+06       0        1
## agegp.Q:tobgp.Q         -5.316e+01  1.890e+06       0        1
## agegp.C:tobgp.Q         -2.083e+01  1.094e+06       0        1
## agegp^4:tobgp.Q         -2.220e+01  9.134e+05       0        1
## agegp^5:tobgp.Q         -1.313e+00  1.626e+05       0        1
## agegp.L:tobgp.C         -2.543e+01  1.210e+06       0        1
## agegp.Q:tobgp.C         -1.423e+01  1.051e+06       0        1
## agegp.C:tobgp.C         -7.456e+00  5.331e+05       0        1
## agegp^4:tobgp.C         -9.394e+00  5.386e+05       0        1
## agegp^5:tobgp.C          2.459e+00  4.054e+04       0        1
## alcgp.L:tobgp.L         -5.552e+01  2.111e+06       0        1
## alcgp.Q:tobgp.L         -4.120e+01  1.271e+06       0        1
## alcgp.C:tobgp.L         -2.234e+01  9.951e+05       0        1
## alcgp.L:tobgp.Q         -1.053e+01  1.190e+06       0        1
## alcgp.Q:tobgp.Q         -3.193e+01  7.778e+05       0        1
## alcgp.C:tobgp.Q         -1.231e+01  4.155e+05       0        1
## alcgp.L:tobgp.C          1.387e+01  4.514e+05       0        1
## alcgp.Q:tobgp.C         -1.718e+01  3.995e+05       0        1
## alcgp.C:tobgp.C          4.239e-01  2.647e+04       0        1
## agegp.L:alcgp.L:tobgp.L -2.840e+02  8.518e+06       0        1
## agegp.Q:alcgp.L:tobgp.L -2.551e+02  7.548e+06       0        1
## agegp.C:alcgp.L:tobgp.L -1.191e+02  4.618e+06       0        1
## agegp^4:alcgp.L:tobgp.L -1.355e+02  3.302e+06       0        1
## agegp^5:alcgp.L:tobgp.L -1.160e+01  8.210e+05       0        1
## agegp.L:alcgp.Q:tobgp.L -1.933e+02  5.800e+06       0        1
## agegp.Q:alcgp.Q:tobgp.L -1.619e+02  5.002e+06       0        1
## agegp.C:alcgp.Q:tobgp.L -7.717e+01  2.661e+06       0        1
## agegp^4:alcgp.Q:tobgp.L -8.845e+01  2.509e+06       0        1
## agegp^5:alcgp.Q:tobgp.L  7.564e+00  2.634e+05       0        1
## agegp.L:alcgp.C:tobgp.L -1.086e+02  4.072e+06       0        1
## agegp.Q:alcgp.C:tobgp.L -9.399e+01  3.573e+06       0        1
## agegp.C:alcgp.C:tobgp.L -4.413e+01  2.055e+06       0        1
## agegp^4:alcgp.C:tobgp.L -4.470e+01  1.524e+06       0        1
## agegp^5:alcgp.C:tobgp.L -3.678e+00  2.280e+05       0        1
## agegp.L:alcgp.L:tobgp.Q -1.033e+02  5.251e+06       0        1
## agegp.Q:alcgp.L:tobgp.Q -8.689e+01  4.573e+06       0        1
## agegp.C:alcgp.L:tobgp.Q -1.386e+01  2.546e+06       0        1
## agegp^4:alcgp.L:tobgp.Q -8.360e+01  2.212e+06       0        1
## agegp^5:alcgp.L:tobgp.Q  1.969e+01  4.519e+05       0        1
## agegp.L:alcgp.Q:tobgp.Q -1.631e+02  3.877e+06       0        1
## agegp.Q:alcgp.Q:tobgp.Q -1.447e+02  3.275e+06       0        1
## agegp.C:alcgp.Q:tobgp.Q -5.394e+01  1.606e+06       0        1
## agegp^4:alcgp.Q:tobgp.Q -7.473e+01  1.773e+06       0        1
## agegp^5:alcgp.Q:tobgp.Q         NA         NA      NA       NA
## agegp.L:alcgp.C:tobgp.Q -4.418e+01  1.984e+06       0        1
## agegp.Q:alcgp.C:tobgp.Q -4.237e+01  1.646e+06       0        1
## agegp.C:alcgp.C:tobgp.Q -2.202e+01  8.027e+05       0        1
## agegp^4:alcgp.C:tobgp.Q -1.989e+01  8.172e+05       0        1
## agegp^5:alcgp.C:tobgp.Q         NA         NA      NA       NA
## agegp.L:alcgp.L:tobgp.C  2.881e+01  2.178e+06       0        1
## agegp.Q:alcgp.L:tobgp.C  3.731e+01  1.871e+06       0        1
## agegp.C:alcgp.L:tobgp.C  1.822e+01  9.043e+05       0        1
## agegp^4:alcgp.L:tobgp.C  1.169e+01  9.950e+05       0        1
## agegp^5:alcgp.L:tobgp.C         NA         NA      NA       NA
## agegp.L:alcgp.Q:tobgp.C -8.620e+01  1.832e+06       0        1
## agegp.Q:alcgp.Q:tobgp.C -6.466e+01  1.511e+06       0        1
## agegp.C:alcgp.Q:tobgp.C -3.978e+01  8.913e+05       0        1
## agegp^4:alcgp.Q:tobgp.C -3.334e+01  8.036e+05       0        1
## agegp^5:alcgp.Q:tobgp.C         NA         NA      NA       NA
## agegp.L:alcgp.C:tobgp.C -2.805e+00  2.214e+05       0        1
## agegp.Q:alcgp.C:tobgp.C         NA         NA      NA       NA
## agegp.C:alcgp.C:tobgp.C         NA         NA      NA       NA
## agegp^4:alcgp.C:tobgp.C         NA         NA      NA       NA
## agegp^5:alcgp.C:tobgp.C         NA         NA      NA       NA
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.2724e+02  on 87  degrees of freedom
## Residual deviance: 3.0119e-10  on  0  degrees of freedom
## AIC: 323.48
##
## Number of Fisher Scoring iterations: 25
```

```
lmods = step(lmod, direction = "both")
```

```
## Start:  AIC=323.48
## cbind(ncases, ncontrols) ~ agegp * alcgp * tobgp

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                       Df Deviance    AIC
## - agegp:alcgp:tobgp 37   16.109 265.59
## <none>                    0.000 323.48

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=265.59
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     agegp:tobgp + alcgp:tobgp

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                       Df Deviance    AIC
## - agegp:tobgp       15   27.146 246.63
## - agegp:alcgp       15   34.364 253.84
## - alcgp:tobgp        9   23.776 255.26
## <none>                   16.109 265.59
## + agegp:alcgp:tobgp 37    0.000 323.48
##
## Step:  AIC=246.63
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     alcgp:tobgp

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                 Df Deviance    AIC
## - alcgp:tobgp  9   33.796 235.28
## - agegp:alcgp 15   47.484 236.96
## <none>             27.146 246.63
## + agegp:tobgp 15   16.109 265.59
##
## Step:  AIC=235.28
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##                 Df Deviance    AIC
## - agegp:alcgp 15   53.973 225.45
## <none>              33.796 235.28
## - tobgp         3   44.151 239.63
## + alcgp:tobgp  9   27.146 246.63
## + agegp:tobgp 15   23.776 255.26
##
## Step:  AIC=225.45
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
##
##                 Df Deviance    AIC
## <none>              53.973 225.45
## - tobgp         3   64.572 230.05
## + agegp:alcgp 15   33.796 235.28
## + alcgp:tobgp  9   47.484 236.96
## + agegp:tobgp 15   41.455 242.94
## - alcgp         3  120.028 285.51
## - agegp         5  131.484 292.96
```

```
# test the deviance
pchisq(lmods$deviance, lmods$df.residual, lower = FALSE)
```

```
## [1] 0.9738352
```

```
df <- esoph %>%
  mutate(proportion=ncases/(ncontrols+ncases)) %>%
  mutate(weight=(ncontrols+ncases))
predprob <- predict(lmods, type = "response")

# Pearson chi-square statistic
px2 <- sum((df$ncases - df$weight*predprob)^2 /
             (df$weight*predprob*(1 - predprob)))
pchisq(px2, lmods$df.residual, lower.tail = FALSE)
```

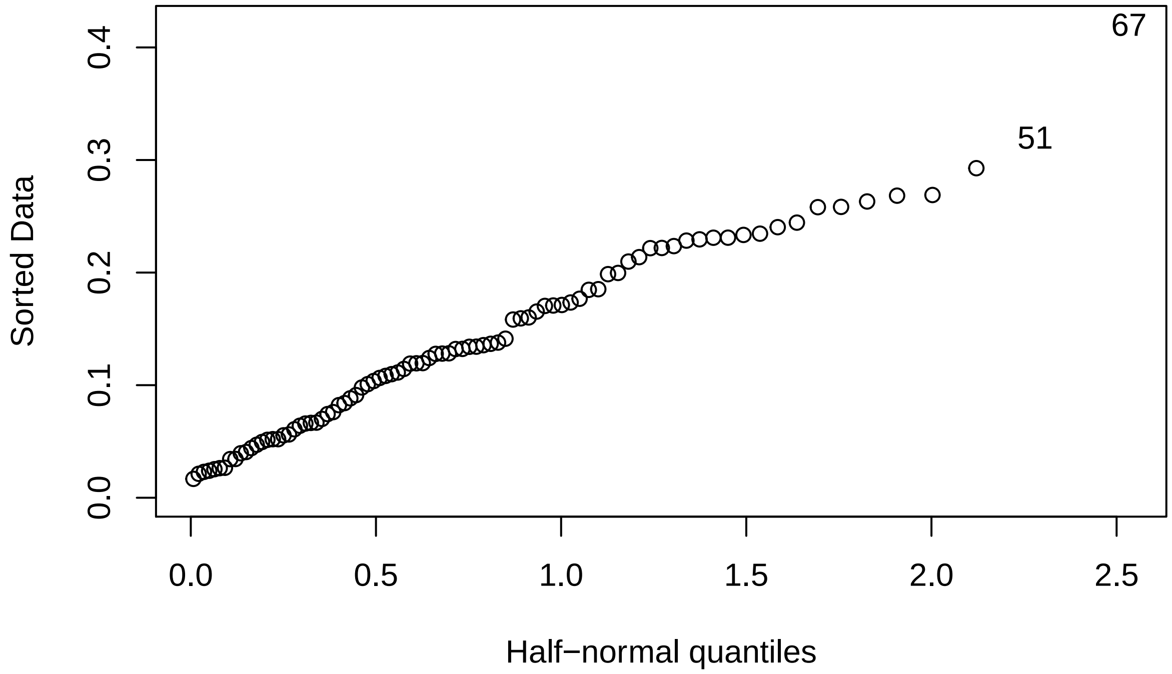
```
## [1] 0.9146142
```

We conducted pearson chi-squre test on the deviance D and Pearson chi-square statistic. The large p-value indicates that the model has an adequate fit.

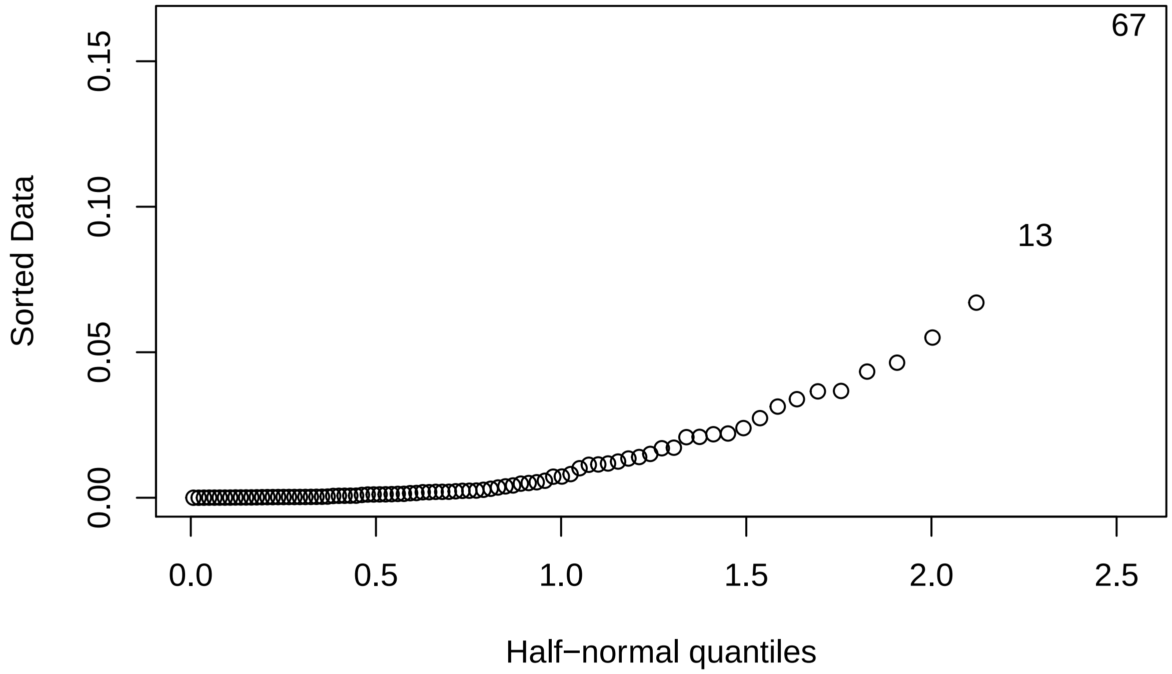
f. Check for outliers in your final model.

Solution:

```
df %>%
  mutate(devres = residuals(lmods, type = "deviance"))%>%
  mutate(linpred = predict(lmods, type = "link")) -> df
halfnorm(hatvalues(lmods))
```



```
halfnorm(cooks.distance(lmods))
```



```
df %>%
  slice(c(13, 51, 67))
```

```
##   agegp alcgp    tobgp ncases ncontrols proportion weight     devres   linpred
## 1 25-34  120+    10-19      1         1  0.5000000      2  2.0642544 -3.454649
## 2 55-64 40-79 0-9g/day      9        40  0.1836735     49 -0.1657969 -1.430876
## 3 65-74 40-79 0-9g/day     17        34  0.3333333     51  1.2146290 -1.062217
```

g. What is the predicted effect of moving one category higher in alcohol consumption?

Solution:

```
lmods %>% summary
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
##     family = binomial, data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6891  -0.5618  -0.2168   0.2314   2.0642
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.77997    0.19796  -8.992  < 2e-16 ***
## agegp.L      3.00534    0.65215   4.608 4.06e-06 ***
## agegp.Q     -1.33787    0.59111  -2.263  0.02362 *
## agegp.C      0.15307    0.44854   0.341  0.73291
## agegp^4      0.06410    0.30881   0.208  0.83556
## agegp^5     -0.19363    0.19537  -0.991  0.32164
## alcgp.L      1.49185    0.19935   7.484 7.23e-14 ***
## alcgp.Q     -0.22663    0.17952  -1.262  0.20680
## alcgp.C      0.25463    0.15906   1.601  0.10942
## tobgp.L      0.59448    0.19422   3.061  0.00221 **
## tobgp.Q      0.06537    0.18811   0.347  0.72823
## tobgp.C      0.15679    0.18658   0.840  0.40071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.241  on 87  degrees of freedom
## Residual deviance:  53.973  on 76  degrees of freedom
## AIC: 225.45
##
## Number of Fisher Scoring iterations: 6
```

h. Compute a 95% confidence interval for this predicted effect.

Solution: